

SENSEVAL-2 日本語辞書タスク

白 井 清 昭[†]

SENSEVAL は語義曖昧性解消を対象としたコンテストである。本論文では、第 2 回 SENSEVAL (SENSEVAL-2) における日本語辞書タスクの概要について報告する。日本語辞書タスクでは、語の意味の区別(曖昧性)を岩波国語辞典によって定義した。参加者には、岩波国語辞典、訓練データ、評価データの 3 つが配布された。訓練データは、3,000 個の新聞記事中の単語に正しい語義を付与したコーパスである。一方評価データは、参加者のシステムが語義を選択すべき単語を含んだ新聞記事である。評価単語の種類は、名詞 50、動詞 50、合わせて 100 個である。また各評価単語毎に 100 ずつ語義を選択するため、評価単語の総数は 10,000 である。正解データは、評価対象となる 10,000 個の単語について、二名の作業者が独立に正しい語義を付与して作成した。この際、二者の語義が一致した割合は 0.863 であり、Cohen の κ は 0.657 であった。また、二者の語義が一致しなかった場合には、第三者が正しい語義を選んだ。日本語辞書タスクには、3 団体 7 システムが参加した。ベースラインシステムのスコア(正解率)が 0.726 であるのに対し、一番成績の良かった参加者のシステムのスコアは 0.786 であった。

キーワード: 語義曖昧性解消、国語辞典、コーパスの注釈付け

SENSEVAL-2 Japanese Dictionary Task

KIYOAKI SHIRAI[†]

SENSEVAL is an evaluation exercise for word sense disambiguation programs. This paper describes a Japanese dictionary task in the second SENSEVAL (SENSEVAL-2). This task defined word senses according to a Japanese dictionary, Iwanami Kokugo Jiten. Three data were distributed to the participants: the Iwanami Kokugo Jiten, the training data and the evaluation data. The training data was a word sense tagged corpus made up of 3,000 newspaper articles, while the evaluation data was newspaper articles containing words of which participants' systems should determine correct word senses. The number of target words was 100, 50 nouns and 50 verbs. One hundred instances of each target word were provided, making for a total of 10,000 instances. For constructing a gold standard data, two annotators chose correct word senses for 10,000 instances separately. The inter-tagger agreement of two annotators was 0.863, while Cohen's κ was 0.657. When word senses selected by two annotators didn't agree, the third annotator chose the correct sense between them. 7 systems of 3 organizations participated in a Japanese dictionary task. The best score achieved by participants' systems was 0.786, while the score of the baseline system was 0.726.

KeyWords: *Word Sense Disambiguation, Japanese Dictionary, Corpus Annotation*

[†] 北陸先端科学技術大学院大学情報科学研究科, School of Information Science, Japan Advanced Institute of Science and Technology

1 はじめに

語義曖昧性解消 (Word Sense Disambiguation, 以下 WSD) は機械翻訳, 情報検索など, 自然言語処理の多くの場面で必要となる基礎技術である (Ide and Veronis 1998). SENSEVAL は WSD のコンテストであり, WSD の共通の評価データを作成し, その上で様々なシステム・手法を比較することによって WSD の研究・技術を向上させることを目的としている. SENSEVAL は過去 2 回行われている. 第 1 回の SENSEVAL (Kilgarriff and Palmer 2000) は 1998 年夏に, 第 2 回の SENSEVAL-2 (Yarowsky 2000) は 2001 年春に行われた. SENSEVAL-2 では, 9 言語を対象に 37 研究グループが参加した. 日本語を対象としたタスクとしては, 辞書タスクと翻訳タスクの 2 つが行われた. 辞書タスクでは語の意味の区別 (曖昧性) を国語辞典によって定義し, 翻訳タスクではこれを訳語選択によって定義した. 本論文は, SENSEVAL-2 の日本語辞書タスクについて, タスクの概要, データ, コンテストの結果について報告する.

まず, 日本語辞書タスクの概要について述べる. SENSEVAL-2 では, タスクを lexical sample task と all words task に大別している. lexical sample task は特定 (数十~数百) の単語だけを WSD の対象とし, all words task では評価テキスト中のすべての単語を対象とする. 日本語辞書タスクは lexical sample task である. 以下, 本論文では, 評価の対象として選ばれた単語を評価単語と呼び, 評価単語の評価データ中での実際の出現を評価インスタンス, または単にインスタンスと呼ぶ. 辞書タスクでは, 単語の語義を岩波国語辞典 (西尾, 岩淵, 水谷 1994) の語義立てによって定義した. 参加者は, テキスト中の評価インスタンスに対して, 該当する語義を岩波国語辞典の語釈の中から選択し, その語釈に対応した ID(以下, 語義 ID) を提出する. 評価テキストは毎日新聞の 1994 年の新聞記事を用いた. 語義を決定する評価単語の数は 100 と設定した. また, 評価単語のそれぞれについて 100 インスタンスずつ語義を決めるとした. すなわち, 評価インスタンスの総数は 10,000 である. 本タスクには 3 団体, 7 システムが参加した.

本論文の構成は以下の通りである. 2 節では, 辞書タスクで用いたデータの概要を述べる. 3 節では, 正解データの作成手順について述べる. また, 正解データを作成する際, 1 つの評価インスタンスに対して二人の作業者が独立に正しい語義を選択したが, そのときの語義の一一致率などについても報告する. 4 節では, 参加者のシステムの概要やスコアなどについて述べ, コンテストの結果に関する簡単な考察を行う. 最後に 5 節では, 本論文のまとめを行う.

2 データ

本節では, 辞書タスクで用いられた 3 つのデータ, 岩波国語辞典, 訓練データ, 評価データについて述べる.

むり【無理】

((名・ダナ)) 理を欠くこと。

- ⑦道理に反すること。「一が通れば道理が引っ込む」「君が怒るのは一
もない(=もっともだ)」。理由が立たないこと。「一な願い」
- ①行きにくいのに、押してすること。「一をして出掛ける」「仕事の一
で病気になる」

図 1 岩波国語辞典の「無理」の語釈文

2.1 岩波国語辞典

1節で述べたように、辞書タスクでは、岩波国語辞典によって語義を定義する。岩波国語辞典の見出しの数は 60,321、語義の総数は 85,870 であり、一見出し当たりの平均語義数は 1.42 である。岩波国語辞典の語釈文の例を図 1 に示す。また、岩波国語辞典では、語義は階層構造を持つ。例えば、図 1 では、「理を欠くこと」という語義が⑦,①の語義の上位にある。階層構造の最大の深さは 3 である。

辞書タスクでは、語義の定義として、形態素解析された岩波国語辞典の語釈文と、それに対する語義 ID が参加者に配布された。なお、語釈文の形態素解析結果は人手修正されている。

2.2 訓練データ

訓練データは、毎日新聞の 1994 年の 3,000 記事を解析したコーパスである。このコーパスに付与されている情報を以下にまとめる。

- 形態素情報(分かち書き、品詞、読み、基本形)
コーパスに含まれる形態素数は 880,000 である。これらは人手修正されている。
- UDC コード
各記事には、テキストの分類カテゴリを表わす指標として、国際十進分類法 (Universal Decimal Classification, UDC) によるコード番号 (情報科学技術協会 1994) が人手によって付与されている。
- 語義情報
各単語には、その単語の意味に該当する語義 ID が付与されている。但し、語義 ID はコーパスの全ての単語ではなく、以下の条件を満たす単語のみに付与されている。
 - 名詞、動詞、形容詞のいずれかである
 - 岩波国語辞典に見出しがある

- 多義である

語義が付与されている形態素の総数は 148,558 である。語義 ID は全て人手によって付与された。また、1 つの単語に語義 ID を付与した人は 1 人である。複数の人が同じ単語に語義 ID を付与し、それらを照合するといった作業は行われていない。

2.3 評価データ

評価データは、評価インスタンスとその正解となる語義 ID を含むテキストである。評価テキストとして毎日新聞の 1994 年の記事を用いた。これらは訓練データの記事とは異なる。評価データに付与されている情報は以下の通りである。

- 形態素情報 (分かち書き、品詞)

これらは自動解析されたものである。訓練データとは異なり、人手による修正はされていない。したがって、訓練データで学習した WSD システムを評価データに適用した際、訓練データと評価データにおける分かち書きや品詞付けの違いによって誤りが生じる可能性がある。本来は評価データの形態素情報も人手修正するべきであったが、今回は準備期間が短かったために断念した。

- UDC コード

訓練データにおける UDC コードと同じ。

- 語義情報 (正解データ)

評価インスタンスには正解となる語義 ID が付与されている。また、訓練データとは異なり、1 つのインスタンスに対して最低 2 人の人が語義 ID を付与している（詳細は 3 節を参照）。もちろん、この情報はコンテストの際には参加者に配布されない。

2.4 付加情報

本節で述べた岩波国語辞典、訓練データ、評価データの付加情報のほとんどは、RWCP によって作成され、1997 年から既に公開されているデータである。訓練データの語義情報については（白井、柏野、橋本、徳永、有田、井佐原、荻野、小船、高橋、長尾、橋田、村田 2001），それ以外の情報については（Hasida, Isahara, Tokunaga, Hashimoto, Ogino, Kashino, Toyoura and Takahashi 1998）を参照していただきたい。これに対し、評価データの語義情報、すなわち正解となる語義 ID のデータは、今回のコンテストのために新たに作成した。3 節では、正解データの作成過程ならびにその概要について述べる。

3 正解データの作成

正解データの作成は以下のように行った。まず評価単語を 100 語選定した。次に、各評価単語毎に 100、合計 10,000 の評価インスタンスを選定した。さらに、各評価インスタンスに対し、のべ二人の作業者が語義 ID を付与した。本節では、正解データ作成の過程、ならびに二者の語義 ID の一致度などについて報告する。

3.1 評価単語、評価インスタンスの選定

評価単語を選定する際には、以下の点を考慮した。

- 評価単語の品詞は名詞または動詞とした。
- 訓練データにおける出現頻度が 50 以上の単語を評価単語とした。
- 訓練データにおける語義の頻度分布のエントロピー $E(w)$ を考慮した。 $E(w)$ の定義を式 (1) に示す。

$$E(w) = - \sum_i p(s_i|w) \log p(s_i|w) \quad (1)$$

式 (1)において、 $P(s_i|w)$ は単語 w の語義が s_i となる確率を表わす。 $E(w)$ の値が大きい単語は、語義の頻度分布が一様であり、語義を決定することが比較的難しい単語であると考えられる。一方、 $E(w)$ の値が小さい単語は、1 つの語義が集中して現われる傾向が強く、語義の決定も比較的易しいと考えられる。

評価単語の選定の際には、 $E(w)$ を WSD の難易度の目安とした。具体的には、以下の 3 つの難易度クラスを設定し、それぞれのクラスから評価単語をまんべんなく選ぶようにした。

- (1) 高難易度の単語クラス $C_{\text{難}}(E(w) \geq 1)$
- (2) 中難易度の単語クラス $C_{\text{中}}(0.5 \leq E(w) < 1)$
- (3) 低難易度の単語クラス $C_{\text{易}}(E(w) < 0.5)$

品詞別、難易度クラス別の評価単語数の内訳を表 1 に示す。また、評価単語の一覧を付録 A に示す。表 1 において、「語義数」は評価単語の岩波国語辞典における語義の数の平均を、「 $E(w)$ 」は評価単語毎に求めた訓練データにおけるエントロピーの平均を表わす。

次に、評価テキストである 1994 年の毎日新聞の記事中から評価インスタンスを選択した。これらの記事には、RWCP によって、形態素情報と UDC コードが付加情報として与えられている。各評価単語毎に、日付の古い記事から順に 100 語を選択し、それらを評価インスタンスとした。ただし、訓練データの記事や UDC コードが付与されていない記事は対象外とした。評価単語は 100 語であるので、評価インスタンスの総数は 10,000 である。また、評価インスタンスが選ばれた記事の総数は 2,130 となった。

表 1 評価単語数の内訳

		C難	C中	C易	計
	単語数	10	20	20	50
名詞	語義数	9.1	3.7	3.3	4.6
	$E(w)$	1.19	0.723	0.248	0.627
動詞	単語数	10	20	20	50
	語義数	18	6.7	5.2	8.3
	$E(w)$	1.77	0.728	0.244	0.743
計	単語数	20	40	40	100
	語義数	14	5.2	4.2	6.4
	$E(w)$	1.48	0.725	0.246	0.685

3.2 語義 ID の付与

10,000語の評価インスタンスに対して、その単語の意味に該当する語義 ID を人手で付与した。語義 ID を付与した作業者は 6 名で、言語学や辞書編纂の知識をある程度持っている人達である。また、本タスクの訓練データは RWCP が作成したコーパスを利用しているが、今回の作業者の中には訓練データへ語義 ID を付与した人も含まれる。その手順を以下にまとめると。

- (1) 二人の作業者が独立に語義 ID を付与する。その際の大まかな指針は以下の通りである。
 - 1つの語義 ID を選択する。複数の語義 ID は選択しない。
 - どの階層の語義 ID を選んでもよい。
 - 岩波国語辞典の語釈の中に該当するものがなければ、UNASSIGNABLE(該当無し)とする。ただし、なるべく UNASSIGNABLE とすることは避け、岩波国語辞典の語釈の中から語義 ID を選択する。
- (2) 二者が選んだ語義 ID が一致していれば、それを正解の語義 ID とする。
- (3) 二者が選んだ語義 ID が一致していないければ、第三者がその中から正しいと思われるものを選択する。ただし、第三者が、二者が選んだ語義 ID 以外の語義 ID が正しいと判断した場合には、三者が選んだ 3 つの語義 ID の全てを正解とする。

語義 ID を選択する際、どの階層の語義 ID を選んでもよいとしたが、階層構造の末端以外の語義 ID が選択されたインスタンスの数は 94 であり、階層の上の語義 ID はあまり選ばれなかった。また、二者の語義 ID が一致せず、第三者も違う語義 ID を選んだインスタンスの数は

28 であり、その全体に対する割合は 0.3% と非常に少なかった。

表 2 は、作業者二人が最初に選んだ語義 ID の一致率を示したものである。評価インスタンス全体における一致率は 86.3% であった。名詞と動詞とで一致率を比較すると、それほど差が見られないことがわかる。また、名詞、動詞ともに、難易度の高いクラスの単語ほど一致率が低くなるが、その傾向は名詞よりも動詞の方が強いことがわかる。

一方、表 3 は評価単語毎に計算した Cohen の κ (Bakeman and Gottman 1997) の平均を示したものである。 κ とは二系列のデータがどの程度一致しているかを測るためによく用いられる統計的尺度であり、式(2)で与えられる。

$$\kappa = \frac{P_o - P_e}{1 - P_e} \quad (2)$$

$$P_o = \frac{\sum_i x_{ii}}{n} \quad (3)$$

$$P_e = \frac{\sum_{i=1}^k x_{+i} x_{i+}}{n^2} \quad (4)$$

式(3)と(4)において、 n はインスタンスの総数を、 x_{ij} は作業者 A が語義 i 、作業者 B が語義 j を与えたインスタンスの数を、 x_{i+}, x_{+i} はそれぞれ作業者 A, B が語義 i を与えたインスタンスの数を表わす。 P_o は二人の作業者が同じ語義を付与した実際の確率であり、 P_e は二人の作業者の語義付与が独立であるときに同じ語義を付与する期待値である。 κ は両者の比から計算され、その値が大きいほど、二者の語義付与が一致していることを示す。その最大値は 1 である。

評価単語 100 語の κ の平均は 0.657 であり、決して大きいとは言えない。このことは、岩波国語辞典の語釈の中から正しい語義を選択する作業は人間でも難しく、付与される語義が人によって揺れやすいことを示唆している。また、表 3 を見ると、表 2 の一致率とは異なり、名詞で難易度クラスが C 中のときの κ の値が不自然に低いことがわかる。これは、一致率と κ が作業者間の語義の一一致度に関して必ずしも同じ傾向を示すわけではないためと考えられる。例えば、100 個の評価インスタンスに対して、作業者 A が語義₁ を 100 回付与し、作業者 B が語義₁ を 99、語義₂ を 1 回付与したとする。このとき、一致率は 0.99 と高いのに対し、 κ は 0 となる。直観的には、 κ の値はもっと大きいと考えられるが、これは統計的に信頼できる κ を求めるのに十分な量のサンプルがなかったためかもしれない。今回の作業では、1 つの評価単語のインスタンスの数は 100 なので、 κ を求める際のサンプル数 n も 100 である。

表 2 作業者の語義 ID の一致率

	C難	C中	C易	計
名詞	0.809	0.786	0.957	0.859
動詞	0.699	0.896	0.922	0.867
計	0.754	0.841	0.939	0.863

表 3 κ の平均

	C難	C中	C易	計
名詞	0.713	0.526	0.655	0.616
動詞	0.605	0.723	0.722	0.698
計	0.659	0.627	0.687	0.657

4 コンテスト

4.1 参加団体

辞書タスクには 3 団体 7 システムが参加した。参加団体とそのシステムの特徴は以下の通りである。いずれのシステムも訓練データを利用した教師あり学習を行っている。

- 通信総研 (CRL)

以下の 4 つのシステムによって回答を提出した。システム名とその概要は以下の通りである (村田, 内山, 内元, 馬, 井佐原 2001)。

- CRL1

分類器としてサポートベクトルマシンを使用したシステム。学習に用いる素性としては、対象語及びその周辺にある単語の表記、品詞、構文情報、意味クラスや UDC コードなどを用いている。

- CRL2

分類器としてシンプルペイズを使用したシステム。学習に用いる素性は CRL1 と同じ。

- CRL3

シンプルペイズとサポートベクトルマシンの混合モデル。個々の対象単語毎に、それぞれの分類器の精度を学習データを用いたクロスバリデーションによって評価し、精度の高い分類器を選択している。

- CRL4
CRL3 と同じような混合モデル。CRL1 と同じ素性を用いたシンプルベイズとサポートベクトルマシン、CRL1 の素性のうち構文素性を使わないシンプルベイズとサポートベクトルマシンの 4 つの分類器を使用している。
- 奈良先端科学技術大学院大学 (NAIST)
以下の 1 つのシステムによって回答を提出した。その概要は以下の通りである (Takamura, Yamada, Kudoh, Yamamoto and Matsumoto 2001)。
 - NAIST
分類器としてサポートベクトルマシンを用いている。学習に用いる素性は、対象語及びその周辺にある単語の表記や品詞の情報などである。さらに、独立成分分析 (Independent Component Analysis, ICA) や主成分分析 (Principle Component Analysis, PCA) といった手法を用いて、素性空間の再構築を行っている。また、複数の素性空間によって学習された分類器を混合している¹。
- 東京工業大学 (TITECH)
以下の 2 つのシステムによって回答を提出した。システム名とその概要は以下の通りである (八木, 野呂, 白井, 徳永, 田中 2001)。
 - TITECH1
分類器として決定リストを用いている。学習に用いる素性は、対象語及びその周辺にある単語の表記、品詞や UDC コードである。また、訓練データの他に、岩波国語辞典の語釈文中の例文からも決定リストの規則を学習している。
 - TITECH2
TITECH1 とほぼ同じであるが、評価データに付与された形態素情報の誤りを自動修正することを試みている。

4.2 評価基準

SENSEVAL-2 では、全ての言語のタスクにおける共通の評価基準として、以下に述べる 3 つの評価基準がある。辞書タスクでも、この評価基準に従ってシステムの評価を行った。

- fine-grained scoring
正解の語義 ID とシステムの語義 ID が完全に一致していれば正解とする。
- coarse-grained scoring
正解の語義 ID とシステムの語義 ID が、語義の階層構造の一番上の層で一致していれば正解とする。

¹ SENSEVAL-2 の参加システムは文献 (Takamura et al. 2001) のシステムと厳密に同じではない。両者の違いは、複数の分類器を混合する際に、前者ではクロスバリデーションによって最も良いと思われる分類器を選択しているのに対し、後者では重み付き多数決によって複数の分類器を混合している。

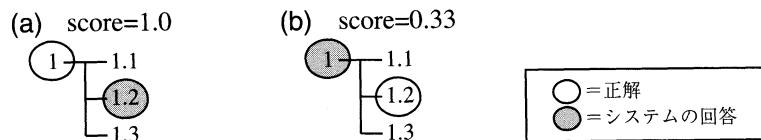


図 2 評価基準 (mixed-grained scoring)

- mixed-grained scoring

正解の語義 ID とシステムの語義 ID が完全に一致していなくても、語義の階層構造に従って部分的にスコアを与える方式で、fine-grained と coarse-grained の中間にあたる。語義の階層構造において、正解の語義 ID がシステムが出力する語義 ID の親であるなら正解とみなす(図 2 (a))。逆に、システムの語義 ID が正解の語義 ID の親であるなら、

$$\frac{1}{\text{システムの語義 ID の子の数}} \quad (5)$$

といった部分的なスコアを与える(図 2 (b))。

参加者は、1つの評価インスタンスに対して複数の語義 ID を返してもよい。また、インスタンスの意味がその語義 ID である確率をつけて返してもよい。確率をつけずに複数の語義 ID を回答した場合には、全ての語義 ID の確率が等しいとして取り扱われる。複数の語義 ID が提出されたときには、各語義 ID の確率に従ってスコアの重み付き平均をとる。また、正解の語義 ID が複数ある場合は、正解の語義 ID 毎にスコアを計算し、その和を全体のスコアとする。

4.3 評価結果と考察

本項では、コンテストの結果とそれに関する考察について述べる。まず、システムの評価結果を図 3 に示す。図 3において、“Baseline”は訓練データにおける最頻出語義を選択したときのスコアを、“Agree”は2人の作業者の語義 ID が一致した割合を示している。参加システムの中で一番スコアが良かったのは CRL4 である。しかし、どのシステムもベースラインを上回り、お互いのスコアの差も 3% 程度で、それほど大きな差は見られなかった。

3つの評価基準によるスコアのうち、coarse-grained score は Baseline も含めてほとんど差はない。また、mixed-grained と fine-grained では、システム間の差に見られる傾向はほとんど同じである。そのため、以後の考察は fine-grained score についてのみ行う。

4.3.1 品詞別に見た評価結果

図 4 は、品詞別に見た各システムのスコア (fine-grained) を示したグラフである。ベースラインを比べると、動詞の方が名詞よりも平均エントロピーが大きい(表 1)にも関わらず、約

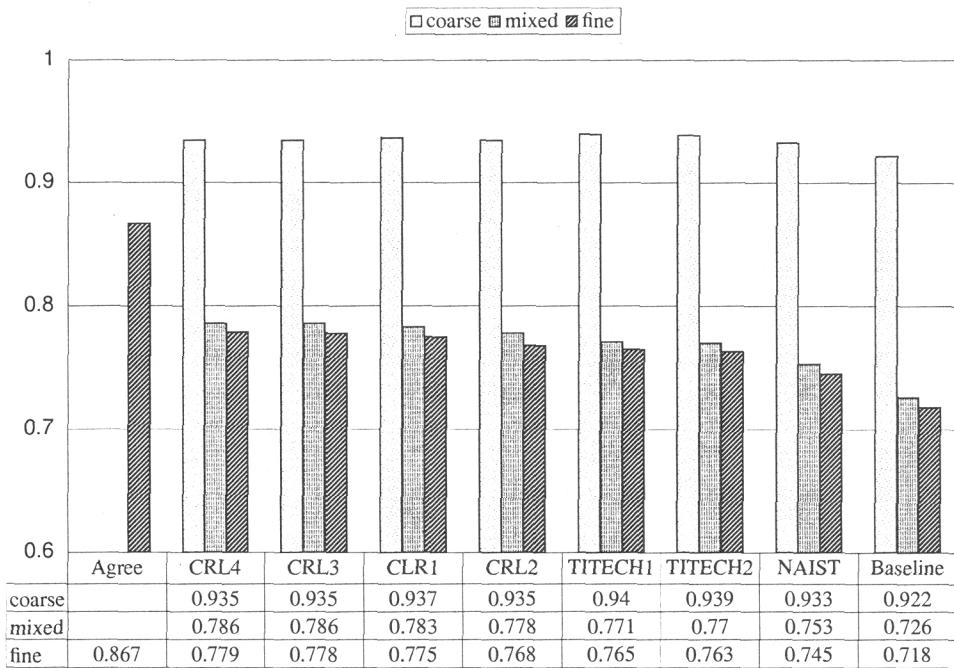


図 3 辞書タスクの結果

3%ほどスコアが高い。これは、特にエントロピーの高い評価単語が動詞にいくつかあり、それらが動詞の平均エントロピーを大きくしているためと考えられる。

参加者のシステムを比べると、名詞のスコアは比較的差が小さいが、動詞のスコアは差が大きい。特に通信総研のシステムは動詞に対するスコアが高く、このことが全体の評価においても他のシステムよりもスコアが高い要因となっている。この原因を明らかにするために、CRL1が正解し NAIST と TITECH2 が不正解であった動詞のインスタンス (139 事例) を抜き出し、どのような動詞に対して CRL のシステムが正しく語義を決めることができるのかについて調査した。通信総研の 4 つのシステムの中から CRL1 を選択したのは、CRL1 が学習アルゴリズムとしてサポートベクトルマシンを採用したシステムであり、同じくサポートベクトルマシンを用いた NAIST と比較するためである。また、東工大の 2 つのシステムの中から TITECH1 を選択したのは、TITECH1 の方が TITECH2 に比べて若干スコアが高いためである。

調査の結果、「描く」「問う」などの動詞について、CRL1 は他のシステムよりも正解率が高いことがわかった。これらの動詞の岩波国語辞典の語釈文と、各システムが出力した語義の頻度を図 5 に示す。しかし、これらの例を見ただけでは、CRL1 が NAIST や TITECH1 に比べて動詞のスコアが高い原因はわからない。原因のひとつとして考えられるのは、CRL1 が NAIST

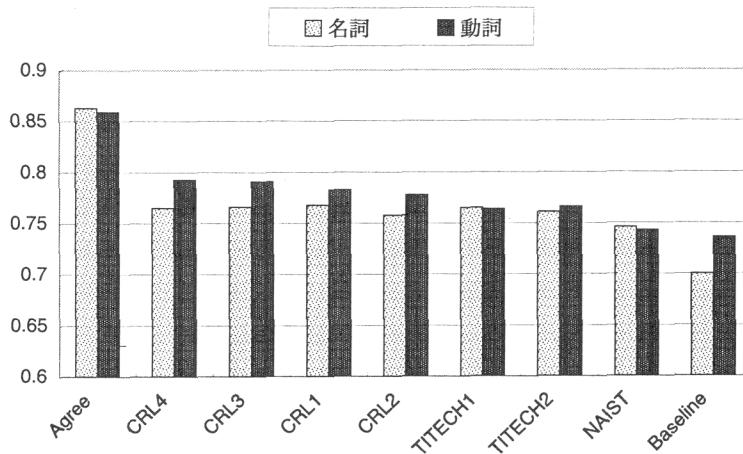


図 4 品詞別スコア (fine-grained)

(a)	「えがく」の語義 (抜粋)	C	N	T	正解
		20	20	20	
① 絵や図をかく。「弧を一いて飛ぶ」		20			○
② 様子を写し出す。表現する。描写する。「情景を一」「勝利を胸に一」					

そのわきで、歴代王たちの肖像を〈描い〉た百メートルにも及ぶ美しい壁画が風雨にさらされている。

(b)	「とう」の語義 (抜粋)	C	N	T	正解
		14	14	14	
[一] 【問う】((五他))		14			○
① わからない事、はっきりしない事を、知らせ(教え)てくれるよう求めめる。問題として出す。「年齢を一わず」(=問題とせず。それで差別しないで)出願できる」					
② 物事の原因、責任の所在、罪を犯した事実などを取り立てて、明らかにするためにただす。「事故の原因を一」「責任を一」					
[二] 【訪う】((五他)) 他人の家や特定の場所を訪問する。おとずれる。たずねる。「恩師を一」「名所旧跡を一」					

交換できる本は汚れのひどくないもので、引き取り価格は一律定価の一〇%。分野は〈問わ〉ず漫画も可。

C,N,T の欄はそれぞれ通信総研、奈良先端大、東工大的システムが該当する語義を出力した頻度を表わす。語釈文の下は、対象インスタンスとそれが現われる新聞記事の例である。対象インスタンスは〈 〉でマークされている。

図 5 CRL1 が正解する動詞の例

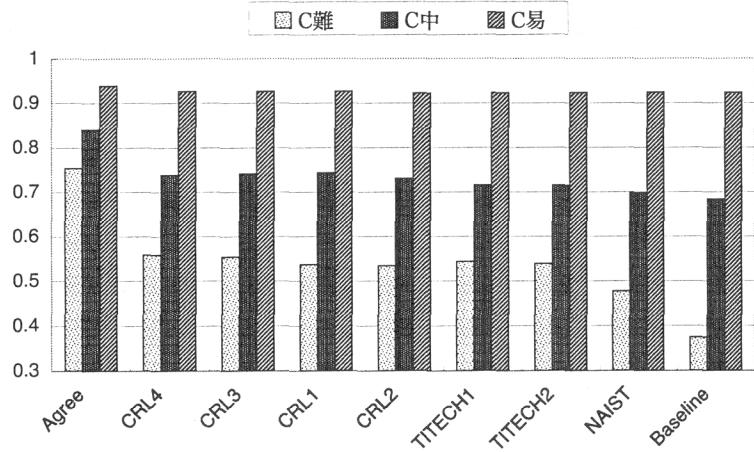


図 6 難易度別スコア (fine-grained)

や TITECH1 と比べて、より多くの素性を用いていることである (4.1項参照)。但し、この推察を裏付けるためには、各システムが個々のインスタンスに対して語義を決める際に手がかりとした素性を明らかにする必要がある。例えば、図 5 に示したインスタンスに対して、CRL1 が NAIST や TITECH2 が考慮していない構文素性などの素性を特に手がかりとしていることが明らかになれば、それらの素性が動詞の語義曖昧性解消に有効であると結論できる。但し、著者は、各システムが語義を決定する際に一番有力な手がかりとした素性に関する情報を持っていないため、上記の考察を具体的に検証することはできなかった。しかし、このように複数の WSD システムの出力を詳細に比較することは、WSD に有効な素性を明らかにし、今後の WSD システムの精度向上につながる可能性がある。

4.3.2 難易度別に見た評価結果

図 6 は、難易度別に見た各システムのスコア (fine-grained) を示したグラフである。クラス C_易 の単語については、ベースラインや作業者的一致率も含めて、各システムのスコアにほとんど差がない。これは、クラス C_易 の単語の語義を決定するタスクが比較的容易であったためと考えられる。これに対し、難易度の高い C_難 や C_中 の単語では、システム間の相違は全体での評価 (図 3) とほぼ同じである。

4.3.3 参加システムの比較

表 4 は、10,000 語の対象インスタンスを (a)3 つの参加者の全てのシステムが正解、(b)1 システムだけが正解、(c)1 システムだけが不正解、(d) 全てのシステムが不正解、の 4 つに分類し、その内訳を調べたものである。通信総研と東工大のシステムとしては一番スコアの良い CRL4

表 4 個々のインスタンスに対する参加システムの比較

	(a)	(b)	(c)	(d)
CRL4	○	○ × ×	× ○ ○	×
NAIST	○	× ○ ×	○ × ○	×
TITECH1	○	× × ○	○ ○ ×	×
	6558	345 283 280	308 501 383	1342

と TITECH1 を選択し、比較の対象とした。また、NAIST は複数の語義を確率付きで回答するシステムであったが、出力された複数の語義の中に正解が含まれていればそのインスタンスに対して正解したとみなすと、NAIST のシステムのパフォーマンスが過大に評価され、システムの公平な比較ができない。そこで、確率の一番大きい語義のみを出力したとみなして他のシステムと比較することにした。ちなみに、確率の一番大きい語義のみを出力したときの NAIST の fine-grained スコアは 0.753 である。

表 4 (b),(c) から、参加者のシステムの回答が完全に一致していない事例の数は 2,100 であることがわかる。これらの事例から、それぞれの WSD システムの特徴を考察することができる。例えば、表 4 (c) の事例は、他のシステムは正しい語義を出力したのに対し、あるシステムだけが正しい語義を出力できなかったことを表わす。付録 B に具体的な事例をいくつか紹介する。これらの事例を調べれば、現在の WSD システムがうまく語義を決定することができない要因を探ることができる。但し、4.3.1 の考察で述べたように、各システムが個々のインスタンスに対して語義を決定する際に一番有力な手がかりとした素性に関する情報が必要である。

また、自然言語処理における様々なタスクにおいて、voting と呼ばれる技術に関する研究が近年盛んに行われている。voting とは、複数のシステムの結果を混合することによりパフォーマンスを向上させる技術で、WSD に応用した研究もいくつか報告されている (Pedersen 2000; Agirre, Rigau, Padró and Atserias 2000; Takamura et al. 2001)。表 4 から、3 つのシステムのいずれかに正解が含まれるインスタンスの割合は 0.865 であることがわかる。これは、3 つのシステムの出力を組み合わせたときに得られるスコアの上限であり、単独のシステムよりも 8% 程度精度が向上することを意味する。したがって、日本語の WSD においても、voting は精度を向上させる技術として有望であろう。

4.3.4 未知の語義

未知の語義とは、ここでは訓練データに 1 回も現われない語義を指す。今回のコンテストでは、未知の語義を正解とするインスタンスの数は 108 であった。未知の語義に対する各システムのスコアを表 5 に示す。各システムは訓練データを用いた機械学習を行っているため、未知

表 5 未知の語義に対するスコア (fine-grained)

CRL1	CRL2	CRL3	CRL4	NAIST	TITECH1	TITECH2
0	0	0	0	0.01	0.648	0.657

「め」の語義 (抜粋)	C	N	T	正解
[一] ((名))				
① 生物の、物を見る働きをする器官。また、その様子・働き。				
⑦ 眼球・視神経から成る器官。	32	49		
① 目①⑦の様子。目つき。				
⑦ 見ること。見えること。また、視力。更に、注意(力)。	37	20		
② 目①⑦に見える姿・様子。				
③ ある物事に出会うこと。経験。体験。また、局面。				
④ 形が目①⑦に似ているもの。「台風の一」				
⋮				
[二] ((接尾))				
① 順序を表す時に添える語。「三番一の問題」	69			☆
⋮				
2年連続13回(目)の優勝を狙う早大を中心に、山梨学院大、中大が追う展開になりそうだ。				

C=CRL4, N=NAIST, T=TITECH1

図 7 未知の語義に対するシステムの出力例

の語義に対するスコアは全体のスコアに比べて著しく劣る。また、参加者のシステムを比較すると、東工大のシステムのスコアが特に高いことがわかる。東工大システムのみが正解した例を図 7 に示す。

図 7 に示したように、[二]①が正解となるインスタンスに対して、TITECH1 は正解と同じ語義を出力するのに対し、CRL4, NAIST は訓練データの頻出語義である [一]①⑦ や [一]①⑦ を出力することがわかった。東工大のシステムが訓練データにない語義を正確に返すのは、語彙文中の例文からも決定リストの規則を学習しているためである。東工大システムの開発者に、「目」の語義を [二]① に決めた決定リストの規則を問い合わせたところ、式 (6) の規則であることがわかった。

対象インスタンスの 1 つ前の単語の品詞が“名詞 接尾 助数詞”かつ 2 つ前の単語の品詞が“名詞 数”なら、語義を [二]①にせよ。 (6)

式(6)の規則は、訓練データの例文ではなく、語義 [二]①の語釈文中の例文「三番一の問題」から学習されたものである。このように、WSD システムを構築する際に複数の知識源を利用する—東工大システムの場合は訓練データ（語義タグ付きコーパス）と辞書の語釈文—は、WSD の精度向上に有効な手段であると考えられる。

なお、訓練データ中に [二]①の語義が現われなかつた理由は以下の通りである。訓練データにおいては、図 7 のような「目」の品詞は“名詞 接尾”になっている。訓練データに語義を付与する際に、接尾語は対象外としたため、これらの単語には語義が付与されていない。ところが、評価データにおいては、RWC の品詞体系の大分類が“名詞”または“動詞”的な単語を対象インスタンスとしたため、品詞が“名詞 接尾”的な単語も語義を決める対象となっている。このため、学習データに含まれない、接尾語としての意味 [二]①を正解とするインスタンスが評価データに頻出した。このような状況は明らかにタスクの設定として不適切である。これは主催者側の過失であり、反省点としたい。

4.3.5 作業者の一致率とシステムのスコア

図 8 は、作業者が付与した語義の一致率を横軸、参加者の 7 システムの平均スコアを縦軸とし、100 個の評価単語の結果をプロットしたグラフである。この図から、作業者の一致率とシステムの平均スコアには正の相関関係があることが読みとれる。しかし、評価単語の中には、作業者の一致率が高いのにも関わらず、システムの平均スコアが低い単語がある。具体的には「開発」「核」「精神」「乗る」「生まれる」「かかる」などである。これらの一連の単語の語義と、参加システムが output した語義の頻度を付録 C に示す。

このような単語は、人間にとっては正しい語義を選択するのは易しいが、現状の WSD システムではうまく語義を決めることができない単語である。したがって、特にこれらの単語について、システムが語義の選択を誤る原因を考察すれば、システムの性能を向上させることができると期待される。

5 おわりに

本論文では、SENSEVAL-2 の日本語辞書タスクの概要について報告した。辞書タスクは、タスク設定自体はオーソドックスなものであるが、日本語を対象とした語義曖昧性解消に関するコンテストとしては始めての試みである。本タスクで用いられた正解データや参加者のシステムの結果は、SENSEVAL-2 のウェブサイト²で公開されている。これらのデータが今後の語

² <http://www.senseval.org/>

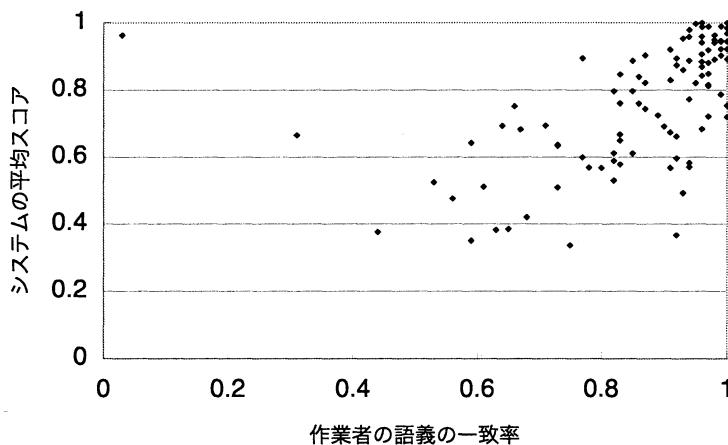


図 8 作業者語義の一致率とシステムの平均スコア (fine-grained)

義曖昧性解消の研究に貢献することを願う。

謝辞

辞書タスクでは、評価テキストとして毎日新聞の新聞記事を利用させていただきました。新聞記事の利用に御協力いただきました毎日新聞社に感謝いたします。また、辞書タスクの運営に数々の助言をいただいた東京工業大学の徳永健伸助教授、東京大学の黒橋禎夫助教授、ならびに正解データを作成して下さった作業者の皆様に深く感謝いたします。査読者の方には、コンテストの結果の考察に関して示唆に富む数多くの御意見をいただきました。厚く御礼申し上げます。

参考文献

- Agirre, E., Rigau, G., Padró, L. and Atserias, J. (2000). "Combining Supervised and Unsupervised Lexical Knowledge Methods for Word Sense Disambiguation." *Computers and the Humanities*, 34 (1,2), pp. 103–108.
- Bakeman, R. and Gottman, J. M. (1997). *Observing Interaction : an Introduction to Sequential Analysis* (2nd edition). Cambridge University Press.
- Hasida, K., Isahara, H., Tokunaga, T., Hashimoto, M., Ogino, S., Kashino, W., Toyoura, J. and Takahashi, H. (1998). "The RWC Text Databases." In *Proceedings of the first International Conference on Language Resources and Evaluation*, pp. 457–462.
- Ide, N. and Veronis, J. (1998). "Introduction to the Special Issue on Word Sense Disambiguation." *Computational Linguistics*, 24 (1), pp. 1–40.

- 情報科学技術協会 (1994). 国際十進分類法 – 日本語中間版 – (第3版). 丸善.
- Kilgarriff, A. and Palmer, M. (2000). "Introduction to the Special Issue on SENSEVAL." *Computers and the Humanities*, 34 (1), pp. 1–13.
- 村田真樹, 内山将夫, 内元清貴, 馬青, 井佐原均 (2001). "SENSEVAL2J 辞書タスクでのCRL の取り組み." 電子情報通信学会技術報告 言語理解とコミュニケーション研究会, pp. 31–38.
- 西尾実, 岩淵悦太郎, 水谷静夫 (1994). 岩波国語辞典 第五版. 岩波書店.
- Pedersen, T. (2000). "A Simple Approach to Building Ensembles of Naive Bayesian Classifiers for Word Sense Disambiguation." In *Proceedings of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics*, pp. 63–69.
- 白井清昭, 柏野和佳子, 橋本三奈子, 徳永健伸, 有田英一, 井佐原均, 萩野紫穂, 小船隆一, 高橋裕信, 長尾確, 橋田浩一, 村田真樹 (2001). "岩波国語辞典を利用した語義タグ付きテキストデータベースの作成." 情報処理学会自然言語処理研究会, 2001 (9), pp. 117–122.
- Takamura, H., Yamada, H., Kudoh, T., Yamamoto, K. and Matsumoto, Y. (2001). "Ensembling based on Feature Space Restructuring with Application to WSD." In *Proceedings of the Natural Language Processing Pacific Rim Symposium*, pp. 41–48.
- 八木豊, 野呂智哉, 白井清昭, 徳永健伸, 田中穂積 (2001). "決定リストを用いた語義曖昧性解消." 電子情報通信学会技術報告 言語理解とコミュニケーション研究会, pp. 47–52.
- Yarowsky, D. (Ed.) (2000). *SENSEVAL-2: Second International Workshop on Evaluating Word Sense Disambiguating Systems*. The Association for Computational Linguistics.

付録

A 評価単語

辞書タスクの評価単語の一覧を以下に示す. また, 日本語タスクには辞書タスクと翻訳タスクの2つがあるが, 両タスクの評価単語は同じものを使用した. ただし, 翻訳タスクの評価単語の数は40である. 下表で*のついた単語は, 翻訳タスクの評価単語でもある.

	C難	C中	C易
名詞	間, 頭, 一般*, 意味*, 姿*, 近く*, 手, 胸*, 目, もの	一方*, 今*, 開発, 関係, 気持ち, 記録*, 国内*, 言葉*, 子供, 午後, 市場, 市民*, 時間, 事業*, 時代*, 情報, 地方, 同日, 場合*, 前*	疑い, 男, 核*, 技術, 現在, 交渉, 社会, 少年, 自分, 精神, 対象, 代表, 中心*, 程度, 電話, 花*, 反対*, 民間, 娘, 問題*
動詞	与える*, 受ける*, かかる, 聞く*, 進む, 出す, 出る*, 取る, 入る, 持つ*	言う*, 訴える, 生まれる, 描く*, 書く*, 決まる, 来る, 超える*, 使う*, 作る*, 伝える*, 出来る, 問う, 残す, 乗る*, 開く, 待つ*, まとめる, 守る*, 見る	思う, 買う*, 変わる, 考える, 決める, 加える, 知る, 進める, 違う, 狙う, 図る*, 話す, 含む, 見せる*, 認める*, 迎える, 求める*, 読む, よる, 分かる

B 1つのシステムだけが不正解となる事例

CRL4, NAIST, TITECH1 のうち, 1つのシステムだけが不正解となった事例を紹介する。

- CRL4 のみが不正解となる事例

「じょうほう」の語義	C	N	T	正解
① ある物事の事情についての知らせ。「海外一」「一を流す」	19	19		☆
② それを通して何らかの知識が得られるようなもの。 $\nabla_{\text{information}}$ の誤語。「データ」が表現の形の面を言うのに対し、内容面を言うことが多い。	19			

「お天気（情報）の大切さを一般の人に理解していただくことが、僕の使命と思っています。…

- NAIST のみが不正解となる事例

「こども」の語義(抜粋)	C	N	T	正解
① 幼い子。児童。	22	22		☆
② 自分のもうけた子。むすこ、むすめ。子。			22	

（子供）のころ、牛肉のすきやきは月に一度ありつけるかどうかのごちそうだった。

- TITECH1 のみが不正解となる事例

「むね」の語義(抜粋)	C	N	T	正解
① 動物の、(体の前面で) 首と腹との間の部分。「一を張る」	12	12		☆
② 胸①の内側に収まっている(と考える)もの。				
⑦ 肺。「一をわざらう」				
① 胃。「一が焼ける」				
⑦ 心臓。「一がどきどきする」				
② 心。「一に迫る」「一に秘める」	12			
同乗の兵庫県西宮市鷺林寺南町、無職、大園輝樹さん(21)が〈胸〉などを強く打つて間もなく死亡。				

C 一致率が高くシステムのスコアが低い単語の例

2人の作業者の一致率が高いのにも関わらず、7つの参加システムの平均スコア(fine-grained)が低い単語の例を以下に示す。以下の表は、作業者の語義付けが一致したインスタンスに対する参加システムの出力語義の頻度分布である。CはCRL4、NはNAIST、TはTITECH1を表わす。

- 「開発」(一致率 0.93, 参加システムの平均スコア 0.493)

2人の作業者がともに正解を①⑦とした場合。

「かいはつ」の語義(抜粋)	C	N	T	正解
① 開きおこすこと。	4			
⑦ (天然資源などを) 人間生活に役立たせること。「電源一」	7	2	5	☆
① 現実化すること。実用化すること。「新製品の一」「研究一」	35	44	41	
② 教育で、問答などを使って自発的にわからせる方法。	「試掘で百本のうち三本も当たれば十分」とされる石油〈開発〉が、なぜ今になって盛り上がってきたのか――。			

- 「核」(一致率 0.97, 参加システムのスコアの平均 0.721)

2人の作業者がともに正解を①とした場合。

「かく」の語義	C	N	T	正解
① 物事の中心（となるもの）。かなめ。「核になる」	1			☆
② 物の中心の部分。「地核・痔核（じかく）」				
⑦ 細胞核。「核膜・核分裂」	21	22	22	
④ 原子核。また、核兵器。「核の持込み」				
③ 草や木の芽ばえるたね。内果皮の硬化したもの。「核果」				
とはいって、このままボスニア情勢を座視していれば、国連を〈核〉とした地域紛争の管理システムの信頼性が決定的打撃を受けかねない。				

- 「乗る」(一致率 0.91, 参加システムのスコアの平均 0.567)

2人の作業者がともに正解を③⑦とした場合。

「のる」の語義(抜粋)	C	N	T	正解
① 運送用の物の上や内部に移る。「馬に一」	19	11	20	
② (持ち上げられて) 物の上に移る。				
⑦ 物に上がる。「台の上に一」	5			
④ 上に置かれる。載「机に一っている本」				
③ 動き・調子によく合う。				
⑦ 勢いがついて物事が進む状態にある。「仕事に気が一らない」	2	4	1	☆
① 他のものの調子にうまく合う。「リズムに一って踊る」				
⑦ 十分によくつく。「あぶらの一った肉」				
⑦ 物事をする仲間・相手になる。「相談に一」				
⑦ 他からのたくらみにまんまと引き込まれる。「計略に一」	1			
④ 伝える手段に託せられる。「電波に一って広まる」。特に、新聞・雑誌・書物に記される。「社会面に一った記事」				
株式会社にして資金を集めブームに〈乗っ〉て事業拡大し、資産を公開して大企業になり結局だれのものだかわからなくなってしまう。				

- 「生まれる」(一致率 1, 参加システムのスコアの平均 0.719)

2人の作業者がともに正解を②とした場合.

「うまれる」の語義	C	N	T	正解
① 母体から子や卵が、その時期が来て、出る。また、卵からかえる。出生する。誕生する。	11	22	20	
② 今までなかったものが出来上がる。	33	22	24	☆
料理は産物で支配され、結果、その国、その土地の伝統的料理が(生まれ)、育ちました。				

略歴

白井 清昭: 1993 年東京工業大学工学部情報工学科卒業, 1998 年同大学院情報理工学研究科博士後期課程修了. 同年同大学院情報理工学研究科計算工学専攻助手. 2001 年北陸先端科学技術大学院大学情報科学研究科助教授, 現在に至る. 博士(工学). 統計的自然言語処理に関する研究に従事. 情報処理学会, 人工知能学会会員.

(2002 年 4 月 30 日 受付)

(2002 年 7 月 28 日 再受付)

(2002 年 9 月 9 日 採録)